

Variance reduced multilevel path simulation: going beyond the complexity ε^{-2}

Denis Belomestny^{a,b,1}, Tigran Nagapetyan^{c,*}

^aDuisburg-Essen University, Duisburg, Germany

^bNational Research University Higher School of Economics, Moscow, Russia

^cWeierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Abstract

In this paper a novel modification of the multilevel Monte Carlo approach, allowing for further significant complexity reduction, is proposed. The idea of the modification is to use the method of control variates to reduce variance at level zero. We show that, under a proper choice of control variates, one can reduce the complexity order of the modified MLMC algorithm down to $\varepsilon^{-2+\delta}$ for any $\delta \in [0, 1)$ with ε being the precision to be achieved. These theoretical results are illustrated by several numerical examples.

Keywords: Control Variates, Wiener Chaos decomposition, Multilevel Monte Carlo, regression

1. Introduction

The multilevel path simulation method introduced in Giles [1] has gained popularity as a complexity reduction tool in recent years. The main advantage of the MLMC methodology is that it can be simply applied to various situations and requires almost no prior knowledge on the path generating process. Any multilevel Monte Carlo (MLMC) algorithm uses a number of levels of resolution, $l = 0, 1, \dots, L$, with $l = 0$ being the coarsest, and $l = L$ being the finest. In the context of a SDE simulation on the interval $[0, T]$, level 0 corresponds to one timestep $\Delta_0 = T$, whereas the level L has 2^L uniform timesteps $\Delta_L = 2^{-L}T$.

Assume that a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t))$ and an \mathbb{R}^m -valued standard Brownian motion (W_t) are given. Let b be a Lipschitz function from \mathbb{R}^d to \mathbb{R}^d , and σ a Lipschitz function from \mathbb{R}^d to $\mathbb{R}^{d \otimes m}$. Consider now a d -dimensional diffusion process (X_t) solving the SDE

$$X_t = b(X_t)dt + \sigma(X_t)dW_t, \quad t \in [0, T] \quad (1)$$

*Corresponding author

Email addresses: belomestny@uni-due.de (Denis Belomestny), nagapetyan@wias-berlin.de (Tigran Nagapetyan)

¹This research was partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

and assume that we would like to estimate the expectation $Y = \mathbb{E}[f(X_T)]$, where f is a Lipschitz function from \mathbb{R}^d to \mathbb{R} . Furthermore, let $X_{l,T}$ be an approximation for X_T using a numerical discretisation with time step Δ_l . The main idea of the multilevel approach pioneered in Giles [1] consists in writing the expectation of the finest approximation $\mathbb{E}[f(X_T^L)]$ as a telescopic sum

$$\mathbb{E}[f(X_{L,T})] = \mathbb{E}[f(X_{0,T})] + \sum_{l=1}^L \mathbb{E}[f(X_{l,T}) - f(X_{l-1,T})] \quad (2)$$

and then applying Monte Carlo to estimate each expectation in the above telescopic sum. One important prerequisite for MLMC to work is that $X_{l,T}$ and $X_{l-1,T}$ are coupled in some way and this can be achieved by using the same discretized trajectories of the underlying diffusion process to construct the consecutive approximations $X_{l,T}$ and $X_{l-1,T}$. The degree of coupling is usually measured in terms of the variance $\text{Var}[f(X_{l,T}) - f(X_{l-1,T})]$. It is shown in Giles [1], that under the conditions:

$$|\mathbb{E}[f(X_{L,T})] - \mathbb{E}[f(X_T)]| \leq c_1 \Delta_L^\alpha, \quad \text{Var}[f(X_{l,T}) - f(X_{l-1,T})] \leq c_2 \Delta_l^\beta, \quad (3)$$

with some $\alpha \geq 1/2$, $\beta > 0$, $c_1 > 0$ and $c_2 > 0$, the computational complexity of the resulting multilevel estimate needed to achieve the accuracy ε (in terms of RMSE) is proportional to

$$\mathcal{C} \asymp \begin{cases} \varepsilon^{-2}, & \beta > 1, \\ \varepsilon^{-2} \log^2(\varepsilon), & \beta = 1, \\ \varepsilon^{-2-(1-\beta)/\alpha}, & 0 < \beta < 1. \end{cases} \quad (4)$$

The above asymptotic estimates show that the reduction of complexity beyond the order ε^{-2} is not possible, doesn't matter how large is $\beta > 1$. This fact motivates a question on the existence of algorithms with complexity order lower than ε^{-2} . Here we give an affirmative answer to this question and propose a modification of the original MLMC algorithm which makes a further complexity reduction possible. Let us note that the existence of such modification does not contradict the general lower bound in [2] as the authors in [2] consider the case of general path dependent functionals of $(X_t)_{t \in [0,T]}$ and we study here the functionals of the form $f(X_T)$ under some additional smoothness assumption on f . Another notable result has been presented in [3], which introduced a deterministic algorithm, which produces a quadrature rule by iteratively applying a Markov transition based on the distribution of a simplified weak Ito-Taylor step. The algorithm, presented there is of completely different nature and also provides a better rates, than the classical Multilevel Monte Carlo algorithm, but its application is limited to one-dimensional case.

The plan of the paper is as follows. In Section 2 we present the variance reduced MLMC algorithm and analyze its complexity. The important ingredient of the new algorithm is a zero level control variate and its choice is discussed in Section 3. In order to compute the control variate in a efficient way, we need MC regression algorithms presented in Section 4. The Section 6.1 is devoted to numerical examples.

2. Main idea

Fix some $0 < L_0 < L$ and consider a random variable M_{L_0} with $E[M_{L_0}] = 0$, then

$$E[f(X_{L,T})] = E[f(X_{L_0,T}) - M_{L_0}] + \sum_{l=L_0+1}^L E[f(X_{l,T}) - f(X_{l-1,T})].$$

As opposite to the representation (2), we start the telescopic sum not at the roughest approximation $\Delta_0 = T$ but at some intermediate one corresponding to Δ_{L_0} . Moreover, at level zero we subtract a zero mean random variable M_{L_0} which can be viewed as a control variate. By fixing a vector of natural numbers $\mathbf{n} = (n_{L_0}, \dots, n_L)$, we can construct a Monte Carlo estimate for $Y = E[f(X_{L,T})]$ via

$$\widehat{Y} \doteq \frac{1}{n_{L_0}} \sum_{i=1}^{n_{L_0}} [f(X_{L_0,T}^{(i)}) - M_{L_0}^{(i)}] + \sum_{l=L_0+1}^L \frac{1}{n_l} \sum_{i=1}^{n_l} [f(X_{l,T}^{(i)}) - f(X_{l-1,T}^{(i)})],$$

where all pairs $(X_{l-1,T}^{(i)}, X_{l,T}^{(i)})$ are independent. Obviously $E[\widehat{Y}] = E[f(X_{L,T})]$ and

$$\begin{aligned} \text{Var}[\widehat{Y}] &= \frac{1}{n_{L_0}} \text{Var}[f(X_{L_0,T}) - M_{L_0}] + \sum_{l=L_0+1}^L \frac{1}{n_l} \text{Var}[f(X_{l,T}) - f(X_{l-1,T})] \\ &\lesssim \frac{1}{n_{L_0}} \text{Var}[f(X_{L_0,T}) - M_{L_0}] + \sum_{l=L_0+1}^L n_l^{-1} \Delta_l^\beta \end{aligned}$$

for some constant $c > 0$, provided the assumption (3) is fulfilled and f is Lipschitz continuous. So we have for the mean square error of \widehat{Y}

$$E[|\widehat{Y} - E[f(X_T)]|^2] \lesssim \Delta_L^{2\alpha} + \frac{1}{n_{L_0}} \text{Var}[f(X_{L_0,T}) - M_{L_0}] + \sum_{l=L_0+1}^L n_l^{-1} \Delta_l^\beta.$$

The complexity analysis can be then conducted by minimising the cost of constructing \widehat{Y} under the condition that RMSE of \widehat{Y} is bounded from above by ε . Let us however first try to understand why the introduction of an additional parameter L_0 and the control variate M_{L_0} may reduce the complexity. If for any l large enough, we could construct a random variable M_l at a cost Δ_l^{-1} in such a way that $\text{Var}[f(X_{l,T}) - M_l] \lesssim \Delta_l^\beta$, then the cost of \widehat{Y} is of order $\sum_{l=L_0}^L n_l \Delta_l^{-1}$. In this case it is optimal to take $L_0 = L$ in order to get for all $\beta > 1$,

$$\text{comp}(\widehat{Y}) \lesssim \varepsilon^{(\beta-1)/\alpha-2}.$$

Note that $(\beta - 1)/\alpha - 2$ can be arbitrary close to 0 if β increases (note that $\alpha = \beta/2$). Of course, the above assumption on M_l is unrealistic since, under this assumption, the cost of computing M_l is basically proportional to the cost of simulating one discretised path of the

process X . In general, the construction of the control variate M_l would require an additional set of discretised “training” paths. Let us now consider a generic family of control variates of the form

$$M_{N,l} \doteq g_{N,l}(X_{l,\cdot}) = g_{N,l}(X_{l,0}, X_{l,\Delta_l}, \dots, X_T), \quad l = 1, \dots, L,$$

for some functions $g_{N,l} : \mathbb{R}^{2^L} \rightarrow \mathbb{R}$, where N stands for the number of discretised “training” paths (with time step of size $\Delta_l = T2^{-l}$) used to construct the function $g_{N,l}$. In particular, we make the following assumptions.

(AP) For any $N \in \mathbb{N}$ and $l \in \{1, \dots, L\}$, a random function $g_{N,l}(z) : \mathbb{R}^{2^L} \rightarrow \mathbb{R}$ is defined on some probability space $(\Omega^N, \mathcal{F}^N, \mathbb{P}^N)$ which is independent of $(\Omega, \mathcal{F}, \mathbb{P})$.

(AC) The cost of constructing the function $g_{N,l}$ is of order $(N/\Delta_l)^{1+2\cdot\kappa_1} \cdot \Delta_l^{2\cdot\kappa_2}$ and the cost of evaluating $g_{N,l}$ on one (new) discretised trajectory is of order $(N/\Delta_l)^{\kappa_1} / \Delta_l^{1-\kappa_2}$ for some $\kappa_1 \geq 0$ and $0 \leq \kappa_2 < 1$.

(AV) It holds

$$\text{Var} [f(X_{l,T}) - M_{N,l}] \lesssim \Delta^{-1+\kappa_2} \left(\frac{\Delta}{N} \right)^{\theta \cdot (1-\kappa_2)} + \Delta_l^\gamma, \quad N \rightarrow \infty$$

for some constants $\theta > 0$ and $\gamma > 0$.

The following theorem gives an upper bound for the complexity of the variance reduced MLMC estimate \hat{Y} under assumption (3), (AC), (AP) and (AV).

Theorem 1. Suppose that the assumptions (3), (AP), (AC) and (AV) hold, and $\beta > 1$, $\theta > \frac{\kappa_1}{1-\kappa_2}$, $(\beta + 1 - 2\kappa_2)(1 + \kappa_1 + \theta(1 - \kappa_2)) + 2\kappa_1 - 2\theta \cdot (1 - \kappa_2) > 0$, and

$$\gamma \geq \beta - \kappa_2 + \frac{(\beta+1-2\kappa_2)\cdot\kappa_1}{\theta \cdot (1-\kappa_2) - \kappa_1}. \quad (5)$$

Then

$$\text{comp}(\hat{Y}) \lesssim \varepsilon^{-2 + \frac{(-2\kappa_1 + 2\theta \cdot (1-\kappa_2)) \cdot (\beta-1)}{(\beta+1-2\kappa_2)(1+\kappa_1+\theta(1-\kappa_2)) + 2\kappa_1 - 2\theta \cdot (1-\kappa_2)}} \quad (6)$$

Discussion. In some cases (see a discussion in Section 4.1), one can set $\kappa_2 = 0$ and $\kappa_1 = 1 - \theta$. Then (5) transforms to

$$\gamma \geq \beta + \frac{(\beta+1)\cdot\kappa_1}{\theta - \kappa_1} = \frac{\theta \cdot \beta + \kappa_1}{\theta - \kappa_1},$$

while (6) has much simpler form

$$\text{comp}(\hat{Y}) \lesssim \varepsilon^{-2 + \frac{(\theta - \kappa_1) \cdot (\beta-1)}{\beta + 2\kappa_1}}. \quad (7)$$

The constraint $(\beta + 1 - 2\kappa_2)(1 + \kappa_1 + \theta(1 - \kappa_2)) + 2\kappa_1 - 2\theta \cdot (1 - \kappa_2) > 0$ is always fulfilled as long as $\kappa_2 \leq \frac{1}{2}$.

3. Construction of control variates

In this section we are going to present a method of constructing control variates satisfying the assumptions (AP), (AC) and (AV). For the ease of notation, we first restrict our analysis to the one-dimensional case and then shortly discuss an extension to the multidimensional case. Our construction of the control variate will be connected to the Wiener Chaos decomposition which we discuss first (see [4] for a detailed exposition). Let $(\phi_i)_{i \geq 1}$ be an orthonormal basis in $L^2(0, T)$. The Wiener chaos of order $p \in \mathbb{N}$ is the L^2 -closure of the vector field spanned by

$$\left\{ \prod_{i \geq 1} H_{p_i} \left(\int_0^T \phi_i(s) dW_s \right) : \sum_{i \geq 1} p_i = p \right\},$$

where H_p is the Hermite polynomial of order p given by the formula

$$H_p(x) \doteq \frac{(-1)^p}{\sqrt{p!}} e^{x^2/2} \frac{d^p}{dx^p} e^{-x^2/2}, \quad p \in \mathbb{N}_0.$$

It is well known that $(H_p)_{p \geq 0}$ is sequence of orthogonal polynomials in $L^2(\mathbb{R}, \mu)$, where μ stands for centered Gaussian measure. We also have

$$\int_{\mathbb{R}} H_p^2(x) \mu(dx) = 1.$$

Every square integrable random variable F , measurable with respect to \mathcal{F}_T , admits the decomposition

$$F = \mathbb{E}[F] + \sum_{k \geq 1} \sum_{|p|=k} c_p \prod_{i \geq 1} H_{p_i} \left(\int_0^T \phi_i(s) dW_s \right) \quad (8)$$

with $p = (p_1, \dots, p_k, \dots) \in \mathbb{N}^{\mathbb{N}}$ and $|p| = \sum_{i \geq 1} p_i$. Taking into account the orthogonality of Hermite polynomials, we derive an expression for the coefficients c_p :

$$c_p = \mathbb{E} \left[F \times \prod_{i \geq 1} H_{p_i} \left(\int_0^T \phi_i(s) dW_s \right) \right].$$

How the representation (8) could be useful for the construction of control variates? In fact, (8) shows that the r. v.

$$M \doteq \sum_{k \geq 1} \sum_{|p|=k} c_p \prod_{i \geq 1} H_{p_i} \left(\int_0^T \phi_i(s) dW_s \right)$$

is a perfect control variate for F . By cutting the above sum at k , we obtain a sequence of control variates

$$M_K \doteq \sum_{k=1}^K \sum_{|p|=k} c_p \prod_{i \geq 1} H_{p_i} \left(\int_0^T \phi_i(s) dW_s \right)$$

which converges to M (note that $E[M_K] = 0$ due to the orthogonality of (ϕ_i) and (H_n)). In the situation where $F = f(X_{\Delta,T})$ and $X_{\Delta,T}$ comes from a discretisation of (1) with a time step $\Delta = T/J$ for some $J \in \mathbb{N}$, it is natural to take $\phi_i(t) \doteq \mathbb{I}(t \in [(i-1)\Delta, i\Delta]) / \sqrt{\Delta}$, $i = 1, \dots, J$. If $X_{\Delta,T}$ is measurable with respect to $\mathcal{G}_J \doteq \sigma(\Delta_1 W, \dots, \Delta_J W)$ with $\Delta_i W \doteq W_{i\Delta} - W_{(i-1)\Delta}$ and $f(X_{\Delta,T}) \in L^2(\mathcal{G}_J, P)$, then we obtain the decomposition

$$f(X_{\Delta,T}) = E[f(X_{\Delta,T})] + \sum_{k \geq 1} \sum_{|p|=k} c_p \prod_{i=1}^J H_{p_i}(\Delta_i W / \sqrt{\Delta}) \quad (9)$$

with $p = (p_1, \dots, p_J) \in \mathbb{N}^J$. The above measurability assumption means that the approximation $X_{\Delta,T}$ involves only uniformly-spaced discrete Brownian increments. This is, for example, the case for the Euler scheme and the Milstein scheme under the commutativity condition. Furthermore, Giles and Szpruch [5] constructed a coupled Milstein scheme that fulfils both the above measurability assumption and the condition (3) with $\beta > 1$. Let us further analyse the decomposition (9). First note that the coefficients in (9) can be computed via

$$c_p = E \left[f(X_{\Delta,T}) \times \prod_{i=1}^J H_{p_i}(\Delta_i W / \sqrt{\Delta}) \right].$$

Moreover, if we cut the summation in (9) at level K , then the control variate

$$M_{K,\Delta} \doteq \sum_{k=1}^K \sum_{|p|=k} c_p \prod_{i=1}^J H_{p_i}(\Delta_i W / \sqrt{\Delta})$$

satisfies

$$\text{Var}(f(X_{\Delta,T}) - M_{K,\Delta}) \lesssim \Delta^{K/2}, \quad \Delta \rightarrow 0,$$

provided $f(X_{\Delta,T})$ is smooth enough in the sense of Malliavin calculus (see [4]). Note that in order to compute all coefficients appearing in $M_{K,\Delta}$ we need $O(J^K)$ operations and this would make the assumption (AC) unfeasible. Can we do something about it? Before we go over to the general case, let us first look at some examples. Consider a simple SDE

$$dX_t = \sigma X_t dW_t, \quad t \in [0, T],$$

with $X_0 = x_0$ and the corresponding Euler approximation

$$Y_{i+1} = Y_i \cdot (1 + \sigma \Delta_i W), \quad i = 1, \dots, J,$$

where $\Delta_i W = W_{i\Delta} - W_{(i-1)\Delta}$ and $\Delta = T/J$. Suppose that we would like to approximate $E[X_T^2]$. It is easy to see that

$$E[Y_J^2] = x_0^2 (1 + \sigma^2 \Delta)^J$$

and using a telescopic sum trick, we derive

$$Y_J^2 - E[Y_J^2] = \sum_{s=1}^J \left(Y_s^2 (1 + \sigma^2 \Delta)^{J-s} - Y_{s-1}^2 (1 + \sigma^2 \Delta)^{J-s+1} \right).$$

Since $\Delta_s W = \frac{Y_s - Y_{s-1}}{\sigma Y_{s-1}}$, we get

$$\begin{aligned} Y_s^2 - Y_{s-1}^2(1 + \Delta\sigma^2) &= 2(Y_{s-1}Y_s - Y_{s-1}^2) + (Y_s - Y_{s-1})^2 - \Delta\sigma^2 Y_{s-1}^2 \\ &= 2\sigma Y_{s-1}^2 \frac{Y_s - Y_{s-1}}{\sigma Y_{s-1}} + \sigma^2 Y_{s-1}^2 \left(\left(\frac{Y_s - Y_{s-1}}{\sigma Y_{s-1}} \right)^2 - \Delta \right) \\ &= 2\sigma Y_{s-1}^2 \Delta_s W + \sigma^2 Y_{s-1}^2 (\Delta_s W^2 - \Delta). \end{aligned}$$

As a result

$$\begin{aligned} Y_J^2 - \mathbb{E}[Y_J^2] &= \sum_{s=1}^J (1 + \sigma^2 \Delta)^{J-s} (2\sigma Y_{s-1}^2 \Delta_s W + \sigma^2 Y_{s-1}^2 (\Delta_s W^2 - \Delta)) \\ &= \sum_{s=1}^J \left(a_{1,s}(Y_{s-1}) H_1 \left(\frac{\Delta_s W}{\sqrt{\Delta}} \right) + a_{2,s}(Y_{s-1}) H_2 \left(\frac{\Delta_s W}{\sqrt{\Delta}} \right) \right) \end{aligned} \quad (10)$$

with $a_{1,s}(y) = 2\sigma \sqrt{\Delta} y^2 (1 + \sigma^2 \Delta)^{J-s}$ and $a_{2,s}(y) = \sqrt{2} \sigma^2 \Delta y^2 (1 + \sigma^2 \Delta)^{J-s}$. Note that the representation (10) has a much simpler form than (9), as it involves only $2 \times J$ coefficients, i.e. the number of coefficients is linear in J . We also see that if the decomposition (10) holds, then the coefficients a_1 and a_2 can be computed via

$$a_{1,j}(Y_{j-1}) = \mathbb{E} \left[Y_j^2 H_1 \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \middle| Y_{j-1} \right], \quad a_{2,j}(Y_{j-1}) = \mathbb{E} \left[Y_j^2 H_2 \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \middle| Y_{j-1} \right].$$

The above example motivates us to look for a similar representation in the general case. In fact, the following general result can be proved.

Theorem 2. Suppose that the approximations $X_{\Delta,j\Delta}$, $j = 1, \dots, J$, satisfy

$$X_{\Delta,j\Delta} = \Phi_j(X_{\Delta,(j-1)\Delta}, \Delta W_j), \quad j = 1, \dots, J, \quad (11)$$

for some functions $\Phi_j : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $\mathbb{E}[|f(X_{\Delta,T})|^2] < \infty$, then

$$f(X_{\Delta,T}) = \mathbb{E}[f(X_{\Delta,T})] + \sum_{k \geq 1} \sum_{j=1}^J a_{k,j}(X_{\Delta,(j-1)\Delta}) H_k(\Delta_j W / \sqrt{\Delta}), \quad (12)$$

where the equality is to be understood in L^2 sense. The coefficients in (12) can be computed via

$$a_{k,j}(x) = \mathbb{E} \left[f(X_{\Delta,T}) H_k \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \middle| X_{\Delta,(j-1)\Delta} = x \right] \quad (13)$$

for $j = 1, \dots, J$ and $k = 1, 2, \dots$

The representation (12) suggests to use the control variates of the form

$$M_{K,\Delta} \doteq \sum_{k=1}^K \sum_{j=1}^J a_{k,j}(X_{\Delta,(j-1)\Delta}) H_k(\Delta_j W / \sqrt{\Delta})$$

for $K = 1, 2, \dots$. The next theorem allows us to assess the quality of these control variates.

Theorem 3. Suppose that f and all functions $\Phi_j(x, w)$, $j = 1, \dots, J$, are $K+1$ times differentiable with at most polynomially growing derivatives (in x and w) that satisfy

$$\sup_x \mathbb{E} \left[\partial_x^k \Phi_j(x, \Delta_j W) \partial_x^l \Phi_j(x, \Delta_j W) \right] = \begin{cases} 1 + O(\Delta), & k = l = 1, \\ O(\Delta), & \max(k, l) > 1, \end{cases} \quad (14)$$

$$\mathbb{E} \left[\left| \partial_w^k \Phi_j(X_{\Delta, (j-1)\Delta}, \Delta_j W) \right|^2 \right] \leq \rho_w, \quad \mathbb{E} \left[\left| f^{(k)}(X_{\Delta, j\Delta}) \right|^2 \right] \leq \rho_f \quad (15)$$

for some constants $\rho_w, \rho_f > 0$ and all $j = 1, \dots, J$, $k = 1, \dots, K+1$. Then

$$\text{Var}[f(X_{\Delta, T}) - M_{K, \Delta}] \leq C_K \Delta^K \quad (16)$$

for some constant C_K not depending on Δ .

Discussion. The number of coefficients $a_{k,j}$ needed to compute $M_{K, \Delta}$ is equal to $J \times K$ which is significantly lower than J^K . Such a reduction of the cost is quite crucial, as it makes the construction of a family of control variates satisfying the assumptions (AP), (AC) and (AV) possible.

Because of the formula (13), all coefficients $(a_{k,j})$ can be computed via a Monte Carlo regression algorithm based on one set of discretised paths of the process X .

Example 4. Consider, for example, the Euler scheme for a one-dimensional SDE of the form (1). Since

$$X_{\Delta, j\Delta} = X_{\Delta, (j-1)\Delta} + b(X_{\Delta, (j-1)\Delta})\Delta + \sigma(X_{\Delta, (j-1)\Delta})\Delta W_j, \quad j = 1, \dots, J,$$

we have

$$\mathbb{E} \left[\left| \partial_x \Phi_j(x, \Delta_j W) \right|^2 \right] = 1 + 2b^{(1)}(x)\Delta + [b^{(1)}(x)]^2 \Delta^2 + [\sigma^{(1)}(x)]^2 \Delta$$

and

$$\begin{aligned} \mathbb{E} \left[\partial_x^k \Phi_j(x, \Delta_j W) \partial_x^l \Phi_j(x, \Delta_j W) \right] &= b^{(k)}(x)b^{(l)}(x)\Delta^2 + \sigma^{(k)}(x)\sigma^{(l)}(x)\Delta \\ \mathbb{E} \left[\partial_x^k \Phi_j(x, \Delta_j W) \partial_x \Phi_j(x, \Delta_j W) \right] &= \\ &= b^{(k)}(x)\Delta + b^{(k)}(x)b^{(1)}(x)\Delta^2 + \sigma^{(k)}(x)\sigma^{(1)}(x)\Delta \end{aligned}$$

for $k, l > 1$. Furthermore

$$\mathbb{E} \left[\left| \partial_w^k \Phi_j(X_{\Delta, (j-1)\Delta}, \Delta_j W) \right|^2 \right] = \begin{cases} \mathbb{E}[\sigma(X_{\Delta, (j-1)\Delta})]^2 & k = 1, \\ 0, & k > 1. \end{cases}$$

Hence the conditions of Theorem 3 are fulfilled if all the derivatives of the coefficient functions b, σ up to order $K+1$ are uniformly bounded and $\mathbb{E}[\sigma(X_{\Delta, (j-1)\Delta})]^2 < \infty$. The latter condition is fulfilled if σ has at most linear growth.

Remark 5. We want to stress out, that the forthcoming presentation of the regression algorithm for coefficients $(a_{k,j})$ estimation is not the only possibility way to proceed with the algorithm. Our presentation is motivated simply by the popularity of global regression in practice.

4. Monte Carlo regression algorithm

Fix a Q -dimensional vector of real-valued functions $\psi = (\psi_1, \dots, \psi_Q)$ on \mathbb{R}^d . Simulate a set of N paths of the Markov chain $X_{\Delta,j\Delta}$, $j = 0, \dots, J$. Let $\alpha_j^k = (\alpha_{j,1}^k, \dots, \alpha_{j,Q}^k)$ be a solution of the following least squares optimization problem:

$$\operatorname{arginf}_{\alpha \in \mathbb{R}^Q} \sum_{i=1}^N \left[\zeta_{j,k}^{(i)} - \alpha_1 \psi_1(X_{\Delta,j\Delta}^{(i)}) - \dots - \alpha_Q \psi_Q(X_{\Delta,j\Delta}^{(i)}) \right]^2 \quad (17)$$

with

$$\zeta_{j,k}^{(i)} \doteq f(X_{\Delta,T}^{(i)}) H_k \left(\frac{\Delta W_j^{(i)}}{\sqrt{\Delta}} \right).$$

Define an estimate for the coefficient function $a_{k,j}$ via

$$\widehat{a}_{k,j}(z) \doteq \alpha_{j,1}^k \psi_1(z) + \dots + \alpha_{j,Q}^k \psi_Q(z), \quad z \in \mathbb{R}^d.$$

It is clear that all estimates $\widehat{a}_{k,j}$ are well defined on the cartesian product of k independent copies of (Ω, \mathcal{F}, P) . The cost ($\text{cost}(\alpha_j^k)$) of computing α_j^k is of order $N \cdot Q^2$, since each α_j^k is of the form $\alpha_j^k = B^{-1}b$ with

$$B_{p,q} \doteq \frac{1}{N} \sum_{i=1}^N \psi_p(X_{\Delta,j\Delta}^{(i)}) \psi_q(X_{\Delta,j\Delta}^{(i)})$$

and

$$b_p \doteq \frac{1}{N} \sum_{i=1}^N \psi_p(X_{\Delta,j\Delta}^{(i)}) \zeta_{k,j}^{(i)},$$

$p, q \in \{1, \dots, Q\}$. Then the complexity of approximating the family of coefficient functions $a_{k,j}$, $k = 1, \dots, K$, $j = 1, \dots, J$, is of order $O(K \cdot N \cdot Q^2 / \Delta)$, provided that $Q < N$.

4.1. Convergence analysis

Define L_2 -norm of a function g with respect to the distribution $P_{\Delta,j}$ of $X_{\Delta,j\Delta}$ via

$$\|g\|_{L^2(P_{\Delta,j})}^2 \doteq \int g^2(x) P_{\Delta,j}(dx).$$

The following result is useful for deriving the convergence rates of $\widehat{a}_{k,j}(z)$ and can be found in [6] (Theorem 11.3).

Theorem 6. Assume that $\sup_x \text{Var}[\zeta_{j,k} | X_{\Delta,j\Delta} = x] \leq \Sigma_k^2 < \infty$ and

$$\|a_{k,j}\|_\infty \leq A_k < \infty, \quad \forall j = 1, \dots, J, \quad k = 1, \dots, K. \quad (18)$$

Then the truncated estimates

$$\bar{a}_{k,j} \doteq T_{A_k} \widehat{a}_{k,j} = \begin{cases} \widehat{a}_{k,j} & |\widehat{a}_{k,j}| \leq A_k, \\ A_k \times \text{sign}(\widehat{a}_{k,j}) & \text{otherwise,} \end{cases}$$

fulfill

$$\mathbb{E}[\|\bar{a}_{k,j} - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta,j-1})}^2] \leq c \max\{\Sigma_k^2, A_k^2\} \frac{Q \log(N)}{N} + 8 \cdot \inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta,j-1})}^2,$$

for some absolute constant $c > 0$, where $\Psi_Q = \text{span}(\psi_q : 1 \leq q \leq Q)$.

Condition (18) is fulfilled, provided f and all functions $\Phi_j(x, w)$, $j = 1, \dots, J$, with all their derivatives of order $K + 1$ are uniformly bounded. In order to get the explicit convergence rates, one need to get an estimate for

$$\inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta,j-1})}^2.$$

This estimate depends on the type of the regression algorithm (global or local), choice of basis functions, the domain, where the solution of SDE lives and many other factors. The following presentation is devoted to a specific choice of global polynomial regression. This choice is motivated by the popularity of this approach in applications.

Fix some $R > 0$. Let $\Psi_{Q_p, R}$ be the set of all piecewise polynomials of degree p w.r.t. an equidistant partition of $[-R, R]$ into Q intervals, in which case $Q_p = O(pQ)$. The following proposition follows from Theorem 6.

Theorem 7. Suppose that f and all functions $\Phi_j(x, w)$, $j = 1, \dots, J$, are $K + p + 1$ times differentiable with bounded derivatives that satisfy the conditions (14)-(15) and

$$\mathbb{E} \left[\left| \partial_w^k \Phi_j(x, \Delta_j W) \right|^2 \right] \leq \rho_w, \quad \mathbb{E} \left[\left| f^{(k)}(X_{\Delta, j\Delta}) \right|^2 \right] \leq \rho_f \quad (19)$$

for all $x \in \mathbb{R}^d$. Moreover, we assume that the approximation scheme has the property

$$\mathbb{P} \left(\sup_{j=1,2,\dots,N} \|X_{\Delta, j\Delta}\| > R \right) \lesssim R^{-\mu}$$

for some $\mu \in (0, \infty)$. Then it holds for the truncated piecewise polynomial estimates $\bar{a}_{k,j}$

$$\mathbb{E}[\|\bar{a}_{k,j} - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta,j-1})}^2] \lesssim \Delta^k \frac{p Q \log(N)}{N} + \left(\frac{R}{Q} \right)^{2p} + \Delta^k \cdot R^{-\mu}. \quad (20)$$

Corollary 8. Set

$$R = \left(\frac{N}{\Delta} \right)^{\frac{\theta}{\theta+\mu}} \cdot \Delta^{\frac{1}{\theta+\mu}} \text{ and } Q \asymp R^\theta \cdot \left(\frac{N}{\Delta \log(N)} \right)^\chi,$$

where $\theta = \frac{2p}{2p+1} = 1 - \chi$, then we get, up to a $\log(N)$ term

$$\mathbb{E}[\|\bar{a}_{k,j} - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta,j-1})}^2] \lesssim \Delta^{\frac{\theta}{\mu+\theta}} \left(\frac{\Delta}{N} \right)^{\frac{\theta\mu}{\mu+\theta}}.$$

Set

$$\widehat{M}_{K,\Delta} = \sum_{k=1}^K \sum_{j=1}^J \bar{a}_{k,j}(X_{\Delta,(j-1)\Delta}) H_k(\Delta_j W / \sqrt{\Delta}),$$

then

$$\begin{aligned} \mathbb{E} \left[\left| \widehat{M}_{K,\Delta} - M_{K,\Delta} \right|^2 \right] &\leq \sum_{k=1}^K \sum_{j=1}^J \mathbb{E} [\| \bar{a}_{k,j} - a_{k,j} \|_{L^2(\mathbb{P}_{\Delta,j-1})}^2] \\ &\lesssim J \cdot \Delta^{\frac{\theta}{\mu+\theta}} \left(\frac{\Delta}{N} \right)^{\frac{\theta\mu}{\mu+\theta}}. \end{aligned}$$

Discussion. Let us now discuss some implications of the above convergence results for complexity analysis of Section 2. First, we see that for any fixed $K > 0$ and $p > 0$, by assuming that f and all functions $\Phi_j(x, w)$, $j = 1, \dots, J$, are $K + p + 1$ times differentiable with at most polynomially growing derivatives with conditions (14)-(15) being satisfied, we can construct the “empirical” control variate $\widehat{M}_{K,\Delta}$ in such a way that

$$\text{Var}[f(X_{\Delta,T}) - \widehat{M}_{K,\Delta}] \lesssim \Delta^K + \Delta^{-\frac{\mu}{\mu+\theta}} \left(\frac{\Delta}{N} \right)^{\frac{\theta\mu}{\mu+\theta}}. \quad (21)$$

The cost of computing $\widehat{M}_{K,\Delta}$, provided $Q \lesssim N$, is of order

$$KNQ_p^2/\Delta = O \left(R^2 \cdot \left(\frac{N}{\Delta} \right)^{1+2x+\delta} \right) = O \left(\left(\frac{N}{\Delta} \right)^{1+2x+\frac{2\theta}{\theta+\mu}+\delta} \cdot \Delta^{\frac{2}{\theta+\mu}} \right)$$

for arbitrary small $\delta > 0$ and the cost of evaluating $\widehat{M}_{K,\Delta}$ on one (new) discretized trajectory is of order

$$KQ_p/\Delta = O \left(\frac{R}{\Delta} \cdot \left(\frac{N}{\Delta} \right)^{x+\delta} \right) = O \left(\left(\frac{N}{\Delta} \right)^{x+\frac{\theta}{\theta+\mu}+\delta} \cdot \Delta^{-\frac{\mu}{\theta+\mu}} \right).$$

As δ can be arbitrarily small, we will set it to 0, to simplify the presentation. The condition $Q < N$ is equivalent to

$$N > \left(\frac{N}{\Delta} \right)^{\frac{\theta}{\theta+\mu}+x} \cdot \Delta^{\frac{1}{\theta+\mu}} \Rightarrow N^{\theta \frac{\mu-x}{\theta+\mu}} > \Delta^{-x \frac{\mu-x}{\theta+\mu}} \Rightarrow N^\theta > \Delta^{-x},$$

which is fulfilled, provided μ is large enough. Now the assumptions (AC) and (AV) hold with

$$\kappa_1 = x + \frac{\theta}{\theta+\mu}, \quad \kappa_2 = \frac{1}{\theta+\mu}.$$

From Theorem 1 we have that the cost of computing the MLMC estimate \widehat{Y} is less than ε^{-2} , provided

$$\mu > \theta \frac{1+2\chi}{\theta-\chi}. \quad (22)$$

Since $p \geq 1$, the constraint (22) is always fulfilled if $\mu > 3\frac{1}{3}$. In the case, when analytical representation for $a_{k,j}(x)$ is available, one can set $\kappa_1 = 0$, $\kappa_2 = 0$, $\theta = 1$ in Theorem 1 to get

$$\text{comp}(\widehat{Y}) \lesssim \varepsilon^{-2+\frac{\beta-1}{\beta}}. \quad (23)$$

Remark 9. A similar analysis can be done for the standard Monte Carlo algorithm (SMC), which, in the case of analytically known $a_{k,j}(x)$, would give us the complexity estimate

$$\text{comp}(\widehat{Y}_{\text{SMC}}) \lesssim \varepsilon^{-1-\frac{1}{\alpha}}, \quad (24)$$

provided $K \geq \alpha$, where α is the weak convergence rate. The advantage of the SMC algorithms is that one can exploit the so-called weak schemes, where simple random variables are used instead of the Brownian motion increments. The extensive analysis of the SMC algorithm with control variates and an extension to weak schemes will be presented in [7].

5. Extension to multidimensional case

The results of Section 3 and Section 4 can be easily extend to the general multidimensional case where $\Delta_i W = (W_{i\Delta}^1 - W_{(i-1)\Delta}^1, \dots, W_{i\Delta}^m - W_{(i-1)\Delta}^m)$. In particular, the representation (9) takes the form

$$f(X_{\Delta,T}) = \mathbb{E}[f(X_{\Delta,T})] + \sum_{k \geq 1} \sum_{|p|=k} c_p \prod_{r=1}^m \prod_{i=1}^J H_{p_{i,r}}(\Delta_i W^r / \sqrt{\Delta}), \quad (25)$$

with a matrix valued multi-index $p = (p_{i,r})$, $i = 1, \dots, J$, $r = 1, \dots, m$, and $|p| = \sum_{i,r} p_{i,r}$. The analogue of the main representation (12) is of the form

$$f(X_{\Delta,T}) = \mathbb{E}[f(X_{\Delta,T})] + \sum_{j=1}^J \sum_{r=1}^m \sum_{1 \leq q_1 < \dots < q_r \leq m} \sum_{p \in \mathbb{N}^r} a_{j,p,q}(X_{\Delta,(j-1)\Delta}) \prod_{i=1}^r H_{p_i}(\Delta_j W^{q_i} / \sqrt{\Delta}).$$

and can be proved along the same lines as (12). In order to compute the corresponding control variate

$$M_{K,\Delta} = \sum_{j=1}^J \sum_{r=1}^m \sum_{1 \leq q_1 < \dots < q_r \leq m} \sum_{p \in \mathbb{N}^r, |p| \leq K} a_{j,p,q}(X_{\Delta,(j-1)\Delta}) \prod_{i=1}^r H_{p_i}(\Delta_j W^{q_i} / \sqrt{\Delta}),$$

we need to determine $O(J \times K^m)$ coefficients, where each coefficient can be computed via

$$a_{j,p,q}(x) = \mathbb{E} \left[f(X_{\Delta,T}) \prod_{i=1}^r H_{p_i}(\Delta_j W^{q_i} / \sqrt{\Delta}) \middle| X_{\Delta,(j-1)\Delta} = x \right].$$

Under smoothness assumption which are similar to ones of Theorem 3, one can show that

$$\text{Var}[f(X_{\Delta,T}) - M_{K,\Delta}] \lesssim \Delta^K, \quad \Delta \rightarrow 0.$$

Another possible representation has the form

$$f(X_{\Delta,T}) = \mathbb{E}[f(X_{\Delta,T})] + \sum_{k \geq 1} \sum_{j=1}^J \sum_{i=1}^m a_{k,j,i} \cdot H_k(\Delta_j W^i / \sqrt{\Delta}), \quad (26)$$

where the coefficients $a_{k,n,j}$ in (12) are given by

$$a_{k,j,i} = \mathbb{E} \left(f(X_{\Delta,T}) H_k(\Delta_j W^i / \sqrt{\Delta}) \middle| X_{\Delta,j-1}, (\Delta_j W^r)_{r=1}^{i-1} \right). \quad (27)$$

In the representation (26), one need to determine only $J \times K \times m$ coefficients, but the conditional expectations in (27) are now functions of a $m + i - 1$ dimensional random vector.

6. Numerical examples

6.1. Euler scheme

Consider again the SDE

$$dX_t = \sigma X_t dW_t, \quad t \in [0, T], \quad (28)$$

with $X_0 = x_0$. Using the Euler approximation

$$Y_{i+1} = Y_i \cdot (1 + \sigma \cdot \Delta_i W), \quad i = 1, \dots, J,$$

with $\Delta_i W = W_{i\Delta} - W_{(i-1)\Delta}$ and $\Delta = T/J$, we are going to approximate

$$\mathbb{E}[X_T^4] = x_0^4 \cdot \exp(6 \cdot \sigma^2 \cdot T).$$

It holds

$$\mathbb{E}[Y_J^4] = x_0^4 \cdot (1 + 6 \cdot \Delta \cdot \sigma^2 + 3 \cdot \sigma^4 \cdot \Delta^2)^N$$

and

$$Y_J^4 - \mathbb{E}[Y_J^4] = \sum_{k=1}^4 \sum_{j=1}^J a_{k,j}(Y_{s-1}) \cdot H_k \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \quad (29)$$

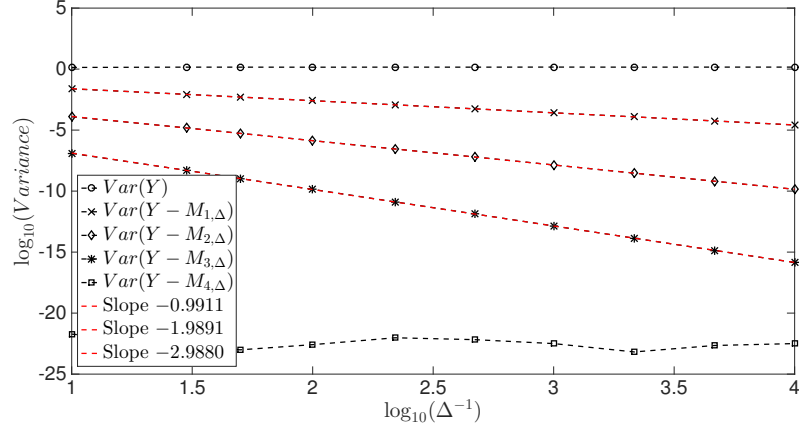


Figure 1: Variance decays with different control variates $M_{K,\Delta}$. Euler scheme.

where

$$\begin{aligned}
 a_{1,j}(y) &= y^4 \cdot 4 \cdot (\sigma \cdot \sqrt{\Delta} + 3 \cdot \sigma^3 \cdot \Delta^{3/2}) \cdot \rho_j, \\
 a_{2,j}(y) &= y^4 \cdot 6\sqrt{2} \cdot (\Delta \cdot \sigma^2 + \sigma^4 \cdot \Delta^2) \cdot \rho_j, \\
 a_{3,j}(y) &= y^4 \cdot 4\sqrt{6} \cdot \Delta^{3/2} \cdot \sigma^3 \cdot \rho_j, \\
 a_{4,j}(y) &= y^4 \cdot \sigma^4 \cdot \Delta^2 \cdot 2\sqrt{6} \cdot \rho_j,
 \end{aligned}$$

with $\rho_j = (1 + 6\sigma^2\Delta + 3\sigma^4 \cdot \Delta^2)^{N-j}$. Let us empirically analyse the performance of the variance reduced MLMC with control variates

$$M_{K,\Delta} = \sum_{k=1}^K \sum_{j=1}^J a_{k,j}(X_{\Delta,(j-1)\Delta}) H_k(\Delta_j W / \sqrt{\Delta}),$$

for $K = 1, 2, 3, 4$. We set $x_0 = 1$, $\sigma = 0.2$, $\Delta_l = 1/J_l$, $l = 1, \dots, 10$, where J_l are logarithmically equally distributed from 10 to 10^4 . The variance decays along with fitted slopes are presented on Figure 1 and are estimated based on $N = 10^6$ samples. One can see, that with control variate $M_{4,\Delta}$ the variance is basically zero, as it is of order 10^{-22} .

6.2. Milstein scheme

Milstein scheme for (28) has the form

$$Y_{i+1} = Y_i \cdot \left(1 + \sigma \cdot \Delta_i W + 0.5 \cdot \sigma^2 \cdot \left((\Delta_i W)^2 - \Delta \right) \right), \quad i = 1, \dots, J. \quad (30)$$

It holds

$$\mathbb{E} [Y_J^4] = x_0^4 \cdot \left(3.75 \cdot \Delta^4 \cdot \sigma^8 + 19 \cdot \Delta^3 \cdot \sigma^6 + 18 \cdot \Delta^2 \cdot \sigma^4 + 6 \cdot \Delta \cdot \sigma^2 + 1 \right)^N$$

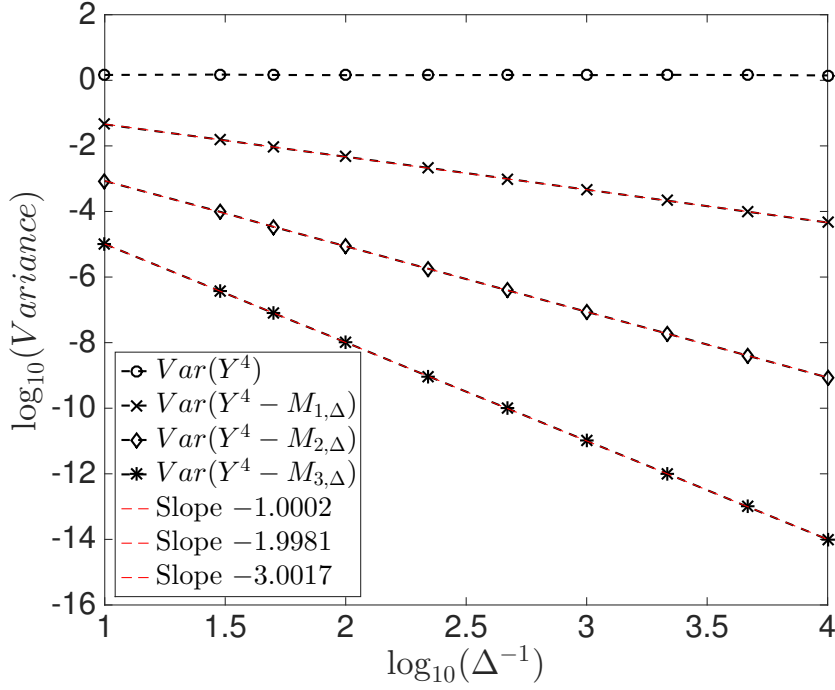


Figure 2: Variance decays with different control variates $M_{K,\Delta}$. Milstein scheme.

and

$$Y_J^4 - \mathbb{E}[Y_J^4] = \sum_{k=1}^8 \sum_{j=1}^J a_{k,j}(Y_{s-1}) \cdot H_k \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \quad (31)$$

The first 4 coefficients in the decomposition (31), have the following form:

$$\begin{aligned} a_{1,j}(y) &= y^4 \cdot \left(34 \cdot \Delta^{7/2} \cdot \sigma^7 + 54 \cdot \Delta^{5/2} \cdot \sigma^5 + 24 \cdot \Delta^{3/2} \cdot \sigma^3 + 4 \cdot \sqrt{\Delta} \cdot \sigma \right) \cdot \rho_j, \\ a_{2,j}(y) &= y^4 \cdot \left(34 \cdot \Delta^4 \cdot \sigma^8 + 132 \cdot \Delta^3 \cdot \sigma^6 + 84 \cdot \Delta^2 \cdot \sigma^4 + 16 \cdot \Delta \cdot \sigma^2 \right) / \sqrt{2} \cdot \rho_j, \\ a_{3,j}(y) &= y^4 \cdot \left(234 \cdot \Delta^{7/2} \cdot \sigma^7 + 252 \cdot \Delta^{5/2} \cdot \sigma^5 + 60 \cdot \Delta^{3/2} \cdot \sigma^3 \right) / \sqrt{6} \cdot \rho_j, \\ a_{4,j}(y) &= y^4 \cdot \left(234 \cdot \Delta^4 \cdot \sigma^8 + 612 \cdot \Delta^3 \cdot \sigma^6 + 204 \cdot \Delta^2 \cdot \sigma^4 \right) / \sqrt{24} \cdot \rho_j, \end{aligned}$$

with $\rho_j = \left(3.75 \cdot \Delta^4 \cdot \sigma^8 + 19 \cdot \Delta^3 \cdot \sigma^6 + 18 \cdot \Delta^2 \cdot \sigma^4 + 6 \cdot \Delta \cdot \sigma^2 + 1 \right)^{N-j}$. Now we estimate the variance decay for different M_{K,Δ_l} , where $K = 1, 2, 3$ and Δ_l are chosen the same as in Subsection 6.1. The variance decay is presented on Figure 2 and is in agreement with Theorem 3.

6.3. Cost - RMSE relation

In order to illustrate the results from Theorems 1 and the estimate (24), we make 100 runs for 5 algorithms for different accuracies.

6.3.1. SMC with Euler scheme

For SMC and SMC with $M_{1,\Delta}$ control variate we use Euler scheme, which has a 1 order weak convergence rate. Moreover, we have $\gamma = 1$, $\theta = 1$, $\kappa_1 = 0$, $\kappa_2 = 0$, as we have all the coefficients in $M_{1,\Delta}$ representation calculated analytically. We run these algorithms for accuracies $\varepsilon_i = 2^{-i}$, $i = 2, 3, \dots, 13$. The results are presented on Figure 3. We can see, that the cost of SMC with $M_{1,\Delta}$ is of order ε^{-2} , which is agreement with (24).

6.3.2. SMC with 2 order weak scheme.

The simplified second order weak Taylor scheme (see p.465 in [8]) for (28) has the form (30), which gives us $\alpha = 2$. The coefficients for $M_{2,\Delta}$ were defined in Subsection 6.2, and again according to (24) we have the cost of order $\varepsilon^{-\frac{3}{2}}$. We run the algorithm for accuracies $\varepsilon_i = 2^{-i}$, $i = 2, 3, \dots, 16$ and the results can be seen on Figure 3. As a finishing remark, we want to state, that any second order algorithm can be used here, if it can be represented in (11) form.

6.3.3. MLMC algorithms

Now in this section, we use scheme (30), but we assume, that we have a weak convergence rate of order 1 for several reasons. First of all, the MLMC complexity does not depend on the weak convergence. Then in general, the Milstein scheme has a first order weak convergence rate. We use accuracies $\varepsilon_i = 2^{-i}$, where $i = 2, 3, \dots, 13$ for the MLMC algorithm and $i = 2, 3, \dots, 16$ for the MLMC with $M_{2,\Delta_{L_0}}$ control variate. The implemented MLMC algorithm corresponds to the standard algorithm from [1], which leads to $\beta = 2$ in the (4) and cost of order ε^{-2} . In the MLMC with $M_{2,\Delta}$ control variate we have, according to Theorem 1 and (23), the overall cost proportional to $\varepsilon^{-2+\frac{\beta-1}{\beta}} = \varepsilon^{-1.5}$. The results can be seen on Figure 3, where we additionally add dotted lines $\text{Cost}^{-\eta}$ with $\eta = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$. The results are in the agreement with Theorems 1 and (24) in the case of analytically calculated coefficients.

6.3.4. MLMC algorithms

Now in this section, we use scheme (30), but we assume, that we have a weak convergence rate of order 1 for several reasons. First of all, the MLMC complexity does not depend on the weak convergence. Then in general, the Milstein scheme has a first order weak convergence rate. We use accuracies $\varepsilon_i = 2^{-i}$, where $i = 2, 3, \dots, 13$ for the MLMC algorithm and $i = 2, 3, \dots, 16$ for the MLMC with $M_{2,\Delta_{L_0}}$ control variate. The implemented MLMC algorithm corresponds to the standard algorithm from [1], which leads to $\beta = 2$ in the (4) and cost of order ε^{-2} . In the MLMC with $M_{2,\Delta}$ control variate we have, according to Theorem 1 and (23), the overall cost proportional to $\varepsilon^{-2+\frac{\beta-1}{\beta}} = \varepsilon^{-1.5}$. The results can be seen on Figure 3, where we additionally add dotted lines $\text{Cost}^{-\eta}$ with $\eta = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$. The results are in the agreement with Theorems 1 and (24) in the case of analytically calculated coefficients.

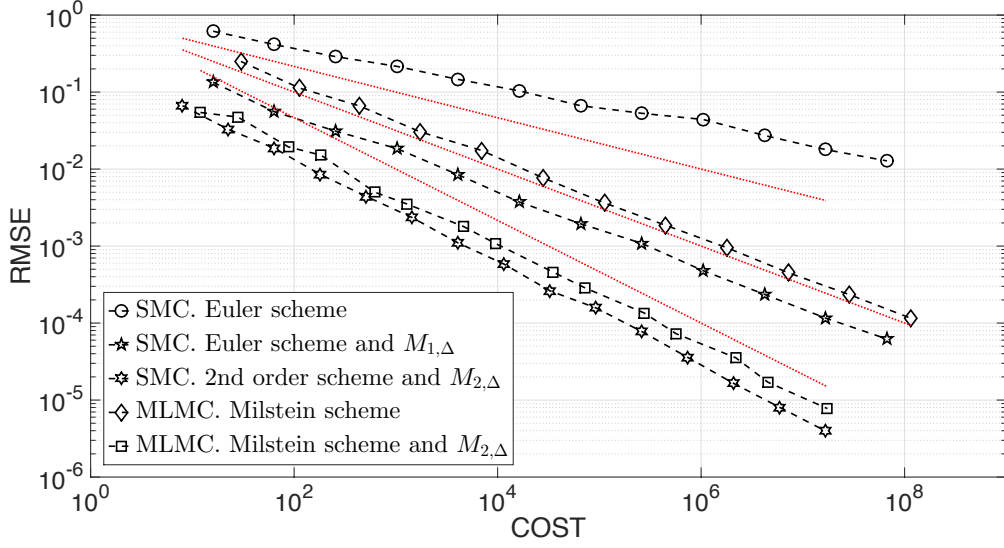


Figure 3: Cost-RMSE relation for the SMC (Euler scheme), SMC(Euler scheme) with $M_{1,\Delta}$ control variate, SMC(second order weak scheme) with $M_{2,\Delta}$ control variate, MLMC (Milstein scheme) and MLMC (Milstein scheme) with $M_{2,\Delta_{L_0}}$ control variate on the level L_0 .

6.4. MLMC with regression

We consider an SDE

$$\begin{aligned} dX_t &= -\sin(X_t) \cdot \cos^3(X_t)dt + \cos^2(X_t)dW_t, \quad t \in [0, T], \\ X_0 &= 0. \end{aligned} \quad (32)$$

This SDE has an exact solution (see [8], p. 121)

$$X_T = \arctan(W_T),$$

and we consider a functional

$$\mathbb{E}f(X_1) = \mathbb{E} \cos(\arctan(W_1)) \approx 0.78964.$$

The considered SDE (32) satisfies requirements in Theorems 6 and 7, moreover, we can take μ in Theorem 7 arbitrarily large, as all polynomial moments exist for the Milstein approximation of SDE (32), so in order to simplify the presentation of our numerical algorithms we set $\mu = \infty$, which means, that the cost estimate has the form (7).

The MLMC algorithm was considered for accuracies $\varepsilon_i = 2^{-i}$, where $i = 2, 3, \dots, 11$, while for the MLMC algorithm with $M_{2,\Delta}$ control variate is considered for $i = 2, 3, \dots, 12$.

We approximate (13) for the $M_{2,\Delta}$ control variate by the global polynomial regression of order $p = 3$, so

$$\theta = 1 - \alpha = \frac{6}{7},$$

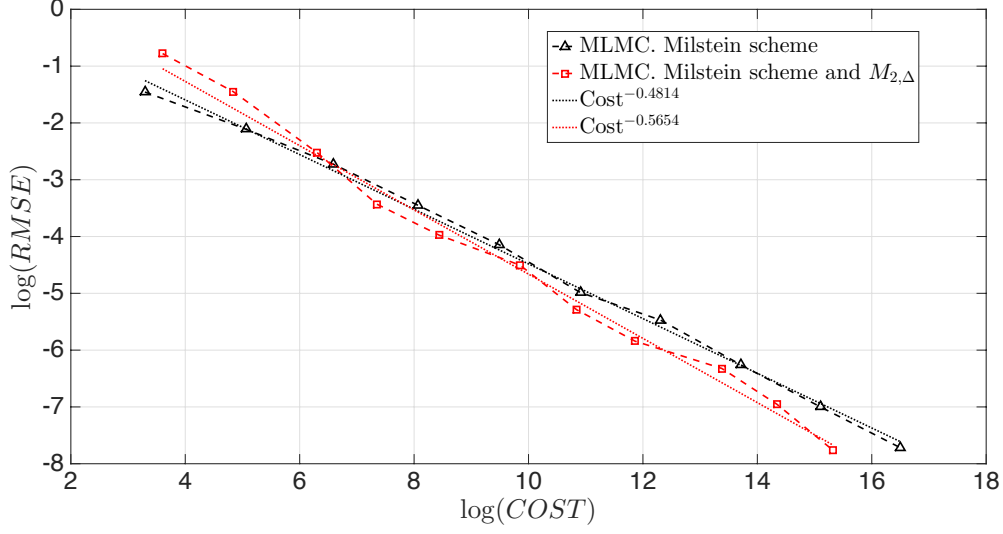


Figure 4: Cost-RMSE relation for the MLMC (Milstein scheme) and MLMC (Milstein scheme) with $M_{2,\Delta_{t_0}}$ control variate with (13) estimated through the regression.

so (7) gives us

$$\mathcal{C} \lesssim \varepsilon^{-2 + \frac{(\theta - \kappa_1) \cdot (\beta - 1)}{\beta + 2\kappa_1}} = \varepsilon^{-\frac{27}{16}} = \varepsilon^{-1.6875}.$$

RMSE for both algorithms is estimated on 100 runs. The results can be seen on Figure 4. The estimated cost are of order 2.0773 for the classical MLMC algorithm (black lines on Figure 4) and 1.7687 for the MLMC algorithm with $M_{2,\Delta}$ control variate (red lines), which is in agreement with the theoretical expectations.

Acknowledgments. Authors are thankful to Prof. Mike Giles and Dr. Sonja Cox for their helpful comments and suggestions on the early versions of this work. Authors are additionally thankful to Dr. Sonja Cox for suggesting multidimensional representation of Theorem 3 in the form (26).

7. Proofs

7.1. Proof of Theorem 1

We write $a \lesssim b$ if there exists a constant $c > 0$ that does not depend on the parameters with respect to which we optimize, such that $a \leq c \cdot b$. Moreover, $a \gtrsim b$ means $b \lesssim a$, and $a \asymp b$ stands for $a \lesssim b$ and $a \gtrsim b$.

The optimization problem has a very simple structure:

$$\mathcal{C} \doteq \left(\frac{N}{\Delta_{L_0}} \right)^{1+2\cdot\kappa_1} \cdot \Delta_{L_0}^{2\kappa_2} + \frac{n_{L_0}}{\Delta_{L_0}} \cdot \left(\frac{N}{\Delta_{L_0}} \right)^{\kappa_1} \cdot \Delta_{L_0}^{\kappa_2} + \sum_{l=L_0+1}^L \frac{n_l}{\Delta_l} \rightarrow \min \quad (33)$$

$$\Delta_L^\alpha \lesssim \varepsilon \quad (34)$$

$$\Delta_{L_0} \geq \Delta_L \quad (35)$$

$$\frac{\Delta_{L_0}^{-1+\kappa_2} \left(\frac{\Delta_{L_0}}{N} \right)^{\theta \cdot (1-\kappa_2)} + \Delta_{L_0}^\gamma}{n_{L_0}} \lesssim \varepsilon^2 \quad (36)$$

$$\sum_{l=L_0+1}^L \frac{\Delta_l^\beta}{n_l} \lesssim \varepsilon^2 \quad (37)$$

$$\Delta_l = 2^{-l}$$

Let us introduce $y = \frac{N}{\Delta_{L_0}}$. Then (33) can be written as

$$\mathcal{C} = y^{1+2\cdot\kappa_1} \cdot \Delta_{L_0}^{2\kappa_2} + \frac{n_{L_0}}{\Delta_{L_0}} \cdot y^{\kappa_1} \cdot \Delta_{L_0}^{\kappa_2} + \sum_{l=L_0+1}^L \frac{n_l}{\Delta_l} \rightarrow \min \quad (38)$$

We set, based on (38),(36) and (37),

$$\begin{aligned} n_{L_0} &\asymp \varepsilon^{-2} \cdot \left(\Delta_{L_0}^{-1+\kappa_2} y^{-\theta \cdot (1-\kappa_2)} + \Delta_{L_0}^\gamma \right) \\ n_l &\asymp \varepsilon^{-2} \cdot \Delta_l^{\frac{\beta+1}{2}} \cdot \sum_{i=L_0+1}^L \Delta_i^{\frac{\beta-1}{2}}, \quad l = L_0 + 1, \dots, L. \end{aligned}$$

We can simplify, provided that $\beta > 1$,

$$\mathcal{C} \asymp y^{1+2\cdot\kappa_1} \cdot \Delta_{L_0}^{2\kappa_2} + \varepsilon^{-2} \cdot y^{\kappa_1} \cdot \Delta_{L_0}^{\kappa_2-1} \cdot \left(\Delta_{L_0}^{-1+\kappa_2} y^{-\theta \cdot (1-\kappa_2)} + \Delta_{L_0}^\gamma \right) + \varepsilon^{-2} \cdot \Delta_{L_0}^{\beta-1}.$$

We introduce an additional constraint

$$\Delta_{L_0}^{-1+\kappa_2} \cdot y^{-\theta \cdot (1-\kappa_2)} > \Delta_{L_0}^\gamma,$$

which will arise later. So now

$$\mathcal{C} \asymp y^{1+2\cdot\kappa_1} \cdot \Delta_{L_0}^{2\kappa_2} + \varepsilon^{-2} \cdot y^{\kappa_1} \cdot \Delta_{L_0}^{2\cdot(\kappa_2-1)} y^{-\theta \cdot (1-\kappa_2)} + \varepsilon^{-2} \cdot \Delta_{L_0}^{\beta-1}.$$

We optimize first w.r.t. y .

$$\begin{aligned} y^{1+2\cdot\kappa_1} \cdot \Delta_{L_0}^{2\kappa_2} &\asymp \varepsilon^{-2} \cdot y^{\kappa_1} \cdot \Delta_{L_0}^{2\cdot(\kappa_2-1)} y^{-\theta \cdot (1-\kappa_2)} \Rightarrow \\ y^{1+\kappa_1+\theta(1-\kappa_2)} &\asymp \varepsilon^{-2} \cdot \Delta_{L_0}^{-2} \Rightarrow N \asymp \Delta_{L_0} \cdot \left(\varepsilon \cdot \Delta_{L_0} \right)^{\frac{-2}{1+\kappa_1+\theta(1-\kappa_2)}}. \end{aligned}$$

The next step is to balance the following terms:

$$\varepsilon^{-2} \cdot y^{\kappa_1 - \theta \cdot (1 - \kappa_2)} \cdot \Delta_{L_0}^{2 \cdot (\kappa_2 - 1)} \asymp \varepsilon^{-2} \cdot \Delta_{L_0}^{\beta - 1} \Rightarrow y^{\kappa_1 - \theta \cdot (1 - \kappa_2)} \asymp \Delta_{L_0}^{\beta + 1 - 2\kappa_2}.$$

Now we introduce $\eta = -2 \frac{\kappa_1 - \theta \cdot (1 - \kappa_2)}{1 + \kappa_1 + \theta(1 - \kappa_2)}$, then

$$(\varepsilon \cdot \Delta_{L_0})^\eta \asymp \Delta_{L_0}^{\beta + 1 - 2\kappa_2} \Rightarrow \Delta_{L_0} = \varepsilon^{\frac{\eta}{\beta + 1 - 2\kappa_2 - \eta}} = \varepsilon^{\frac{-2\kappa_1 + 2\theta \cdot (1 - \kappa_2)}{(\beta + 1 - 2\kappa_2)(1 + \kappa_1 + \theta(1 - \kappa_2)) + 2\kappa_1 - 2\theta \cdot (1 - \kappa_2)}}.$$

And the final answer is

$$\mathcal{C} \asymp \varepsilon^{-2 + \frac{(-2\kappa_1 + 2\theta \cdot (1 - \kappa_2)) \cdot (\beta - 1)}{(\beta + 1 - 2\kappa_2)(1 + \kappa_1 + \theta(1 - \kappa_2)) + 2\kappa_1 - 2\theta \cdot (1 - \kappa_2)}}.$$

Now let us recall the constraint $\Delta_{L_0}^{-1 + \kappa_2} \cdot y^{-\theta \cdot (1 - \kappa_2)} > \Delta_{L_0}^\gamma$, which leads to the following calculations:

$$\begin{aligned} \Delta_{L_0}^{-1 + \kappa_2} \cdot y^{-\theta \cdot (1 - \kappa_2)} &= \Delta_{L_0}^{-1 + \kappa_2 + \frac{(\beta + 1 - 2\kappa_2) \cdot \theta \cdot (1 - \kappa_2)}{\theta \cdot (1 - \kappa_2) - \kappa_1}} = \Delta_{L_0}^{\beta - \kappa_2 + \frac{(\beta + 1 - 2\kappa_2) \cdot \kappa_1}{\theta \cdot (1 - \kappa_2) - \kappa_1}} \geq \Delta_{L_0}^{-\gamma} \Rightarrow \\ \gamma &\geq \beta - \kappa_2 + \frac{(\beta + 1 - 2\kappa_2) \cdot \kappa_1}{\theta \cdot (1 - \kappa_2) - \kappa_1} \end{aligned}$$

7.2. Proof of Theorem 3

First note that for any function $g \in C_{\text{pol}}^k(\mathbb{R})$ and any $j < k$ the integration by parts gives

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(y) H_k(y) e^{-y^2/2} dy &= \frac{(-1)^k}{\sqrt{2\pi k!}} \int_{-\infty}^{\infty} g(y) \frac{d^k}{dy^k} e^{-y^2/2} dy \\ &= \sqrt{\frac{(k-j)!}{2\pi k!}} \int_{-\infty}^{\infty} g^{(j)}(y) H_{k-j}(y) e^{-y^2/2} dy. \end{aligned}$$

Next, for all $j \in \{1, \dots, N\}$ we have a representation

$$X_{\Delta, T} = f(G_j(X_{\Delta, (j-1)\Delta}, \Delta W_j, \dots, \Delta W_J))$$

with a function $G_j : \mathbb{R}^{J-j+2} \rightarrow \mathbb{R}$ defined as

$$G_j(x, \Delta W_j, \dots, \Delta W_J) = \Phi_J(\cdot, \Delta W_J) \circ \Phi_{J-1}(\cdot, \Delta W_{J-1}) \circ \dots \circ \Phi_j(x, \Delta W_j).$$

Hence the integration by parts gives for all $k \in \mathbb{N}$, $j \in \{1, \dots, J\}$, $k' \leq k$

$$\begin{aligned}
a_{k,j}(x) &= \frac{1}{(2\pi)^{(J-j+1)/2}} \int \dots \int f(G_j(x, \sqrt{\Delta}y_j, \dots, \sqrt{\Delta}y_J)) \\
&\quad \times H_k(y_j) e^{-\sum_{i=j}^J y_i^2/2} dy_j \dots dy_J \\
&= \frac{\sqrt{(k-k')!} \Delta^{k'/2}}{(2\pi)^{(J-j+1)/2} \sqrt{k!}} \int \dots \int \frac{\partial^{k'} f(G_j(x, z_j, \sqrt{\Delta}y_{j+1}, \dots, \sqrt{\Delta}y_J))}{\partial z_j^{k'}} \Big|_{z_j=\sqrt{\Delta}y_j} \\
&\quad \times e^{-\sum_{i=j}^J y_i^2/2} H_{k-k'}(y_j) dy_j \dots dy_J \\
&= \frac{\sqrt{(k-k')!} \Delta^{k'/2}}{\sqrt{k!}} \mathbb{E}[\partial_{\Delta_j W}^k f(G_j(x, \Delta_j W, \dots, \Delta_j W)) H_{k-k'}(\Delta_j W)] \quad (39) \\
&= \frac{\sqrt{(k-k')!} \Delta^{k'/2}}{\sqrt{k!}} \mathbb{E}[\partial_{\Delta_j W}^k f(X_{\Delta,T}) H_{k-k'}(\Delta_j W) | X_{\Delta,(j-1)\Delta} = x].
\end{aligned}$$

Then using again integration by parts and the orthogonality of Hermite polynomials, we arrive at

$$\begin{aligned}
\text{Var}[f(X_{\Delta,T}) - M_{K,\Delta}] &\leq \mathbb{E} \left[\left| \sum_{k=K+1}^{\infty} \sum_{j=1}^J a_{k,j}(X_{\Delta,(j-1)\Delta}) H_k(\Delta_j W / \sqrt{\Delta}) \right|^2 \right] \\
&\leq \Delta^K \left[\sum_{m=0}^{\infty} \sum_{j=1}^J \frac{\Delta m!}{(m+K+1)!} \times \right. \\
&\quad \left. \mathbb{E} \left[\mathbb{E}[\partial_{\Delta_j W}^{K+1} f(X_{\Delta,T}) H_m(\Delta_j W / \sqrt{\Delta}) | X_{\Delta,(j-1)\Delta}]^2 \right] \right] \\
&\leq \Delta^K \left[\sum_{m=1}^{\infty} \frac{1}{m^{K+1}} \sum_{j=1}^J \Delta \cdot \mathbb{E} \left[\partial_{\Delta_j W}^{K+1} f(X_{\Delta,T}) \right]^2 \right].
\end{aligned}$$

By di Bruno's formula, we have for $k > j$ and $s = 1, \dots, K$,

$$\begin{aligned}
\partial_{\Delta_j W}^s X_{\Delta,k\Delta} &= \sum_m c_m \partial_x^{|m|} \Phi_k(X_{\Delta,(k-1)\Delta}, \Delta_k W) \prod_{i=1}^s [\partial_{\Delta_j W}^i X_{\Delta,(k-1)\Delta}]^{m_i} \\
&= \partial_x \Phi_k(X_{\Delta,(k-1)\Delta}, \Delta_k W) \partial_{\Delta_j W}^s X_{\Delta,(k-1)\Delta} + \\
&\quad + \sum_{|m|>1} c_m \partial_x^{|m|} \Phi_k(X_{\Delta,(k-1)\Delta}, \Delta_k W) \prod_{i=1}^s [\partial_{\Delta_j W}^i X_{\Delta,(k-1)\Delta}]^{m_i}
\end{aligned}$$

with $c_m = s! \left[\prod_{i=1}^s m_i! (i!)^{m_i} \right]^{-1}$, where the summation is over all multi-indexes $m = (m_1, \dots, m_s)$ satisfying $\sum_{i=1}^s i m_i = s$. Due to the Hölder inequality

$$\mathbb{E} \left[\left| \prod_{i=1}^s Z_i \right| \right] \leq \prod_{i=1}^s \mathbb{E} [|Z_i|^{p_i}]^{1/p_i} \quad \text{with} \quad \sum_{i=1}^s \frac{1}{p_i} = 1$$

we get by setting $1/p_i = im_i/s$:

$$\mathbb{E} \left\{ \prod_{i=1}^s \left[\partial_{\Delta_j W}^i X_{\Delta, (k-1)\Delta} \right]^{2m_i} \right\} \leq \prod_{i=1}^s \left[\mathbb{E} \left\{ \left| \partial_{\Delta_j W}^i X_{\Delta, (k-1)\Delta} \right|^{2s/i} \right\} \right]^{im_i/s}$$

Using the interpolation inequalities for weighted Sobolev spaces (see e.g. [9]) we derive

$$\mathbb{E} \left\{ \left| \partial_{\Delta_j W}^i X_{\Delta, (k-1)\Delta} \right|^{2s/i} \right\} \leq c_1 \mathbb{E} \left\{ \left| \partial_{\Delta_j W}^s X_{\Delta, (k-1)\Delta} \right|^2 \right\}$$

for some constant c_1 which may depend on s . As a result

$$\mathbb{E} \left[\left| \partial_{\Delta_j W}^s X_{\Delta, k\Delta} \right|^2 \right] \leq (1 + c_2 \Delta)^{k-j+1} \rho_w \leq \exp(c_2) \rho_w$$

and using again di Bruno's formula

$$\mathbb{E} \left[\left| \partial_{\Delta_j W}^s f(X_{\Delta, k\Delta}) \right|^2 \right] \leq c_3$$

for $c_3 = c_3(s)$ not depending on j and k .

7.3. Proof of Theorem 7

In the proof of Theorem 3 we derived the representation

$$a_{k,j}(x) = \frac{\Delta^{k/2}}{\sqrt{k!}} \mathbb{E} \left[\partial_j^k f(X_{\Delta, T}) \middle| X_{\Delta, (j-1)\Delta} = x \right]. \quad (40)$$

From the (39) from Theorem 3, one has

$$|a_{k,j}(x)| \leq c \cdot \left(\mathbb{E} \left[\left(\partial_{\Delta_j W}^k f(G_j(x, \Delta_j W, \dots, \Delta_j W)) \right)^2 \right] \right)^{1/2} \cdot \left(\mathbb{E} \left[\left(H_{k-k'}(\Delta_j W) \right)^2 \right] \right)^{1/2},$$

so with (15) one can see that all conditional expectations in (40) are uniformly bounded in x . Hence we can take $A_k = c_1 \Delta^{k/2} / \sqrt{k!}$ in Theorem 6. Analogously due to

$$\text{Var} \left[f(X_{\Delta, T}) H_k \left(\frac{\Delta_j W}{\sqrt{\Delta}} \right) \middle| X_{\Delta, (j-1)\Delta} = x \right] \lesssim \Delta^{2k} / k!$$

we can set $\Sigma_k^2 = c_2 \Delta^{2k} / k!$. Finally we need to estimate

$$\inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta, j-1})}^2.$$

If the basis functions $\psi_q(x)$ are chosen to be zero outside domain $\{x : \|x\| \geq R\}$, then

$$\inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2(\mathbb{P}_{\Delta, j-1})}^2 \leq \inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2_{\|x\| \leq R}(\mathbb{P}_{\Delta, j-1})}^2 + 4 \cdot A_k^2 \cdot \mathbb{P}(\|X_{\Delta, (j-1)\Delta}\| > R).$$

According to the Corollary 11.2 from [6], we have

$$\inf_{q \in \Psi_Q} \|q - a_{k,j}\|_{L^2_{\|x\| \leq R}(\mathbb{P}_{\Delta, j-1})}^2 \leq \left(\frac{R}{Q} \right)^{2p}$$

Now Theorem 6 implies (20).

References

- [1] M. B. Giles, Multilevel Monte Carlo path simulation, *Operations Research* 56 (3) (2008) 607–617.
- [2] J. Creutzig, S. Dereich, T. Müller-Gronbach, K. Ritter, Infinite-dimensional quadrature and approximation of distributions, *Foundations of Computational Mathematics* 9 (4) (2009) 391–429.
- [3] T. Müller-Gronbach, L. Yaroslavl'tseva, Deterministic quadrature formulas for SDEs based on simplified weak Itô-Taylor steps, DFG SPP 1324 (Preprint 167).
- [4] D. Nualart, *The Malliavin calculus and related topics*, Springer, 2006.
- [5] M. B. Giles, L. Szpruch, Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation, *Annals of Applied Probability* 24 (4) (2014) 1585–1620.
- [6] L. Györfi, A. Krzyżak, *A distribution-free theory of nonparametric regression*, Springer, 2002.
- [7] D. Belomestny, S. Haefner, T. Nagapetyan, M. Urusov, Weak Monte Carlo path generation method with optimal control variates., in preparation (2014).
- [8] P. Kloeden, E. Platen, *Numerical solution of stochastic differential equations*, Vol. 23, Springer, 1992.
- [9] S. Boccia, L. Caso, Interpolation inequalities in weighted Sobolev spaces, *Journal of Mathematical Inequalities* 2 (3) (2008) 309–322.